[illegible]

NAME OF AUTHOR [illegible]

TITLE OF THESIS [illegible]

DEGREE FOR [illegible]

YEAR THIS [illegible]

[illegible]

THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR ..... Mikelis G. Bickis .....

TITLE OF THESIS ..... Information and Markov processes .....

DEGREE FOR WHICH THESIS WAS PRESENTED ..... Master of Science .....

YEAR THIS DEGREE GRANTED ..... Fall - 1973 .....

Permission is hereby granted to THE UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

THE UNIVERSITY OF ALBERTA

INFORMATION AND MARKOV PROCESSES

BY

© MIKELIS G. BICKIS

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES & RESEARCH

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

OF MASTER OF SCIENCE

IN

STATISTICS

DEPARTMENT OF MATHEMATICS

EDMONTON, ALBERTA

FALL, 1973

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled ....

INFORMATION AND MARKOV PROCESSES

submitted by Mikelis G. BICKIS

in partial fulfilment of the requirements for the degree of

Master of Science...........

Dedicated to my parents.

ABSTRACT

Several concepts of information can be defined on algebras of events, and can be related to probabilities. Two very useful concepts are entropy and discrimination information, with applications to communication theory and statistical inference, respectively. Conditional information can also be defined, given an event, or sub-algebra of events.

A historical summary of information theory is given in Chapter II, which includes several generalizations of the information concept.

The properties of conditional information are employed in reaching a general result concerning the information in a Markov process.

ACKNOWLEDGEMENT

I want to thank Dr. S. Ghurye, my supervisor, for his understanding and his patience. I want to thank my friends and colleagues, for fruitful discussions, and their encouragement. And I want to thank my wife, Jane, without whose moral support the task would have been far more difficult.

I would like to acknowledge the financial support of the University of Alberta and the National Research Council of Canada.

TABLE OF CONTENTS

Page

CHAPTER I: THE CONCEPT OF INFORMATION

CHAPTER II: A BRIEF HISTORY

CHAPTER III: INFORMATION IN MARKOV PROCESSES

INTRODUCTION

In this dissertation, we will be examining concepts of information and probability, and some applications. Working from an intuitive basis, we will consider how the two concepts are related, and will look at how concepts of information can be utilized in communication theory, statistical inference, and Markov processes.

In chapter I, we introduce measures of probability and information on an algebra of events. Some properties of functions defined on a family of probability algebras are examined. Two such functions are entropy and discrimination information and their characteristics are listed. The contents of this chapter is a reformulation and unification of diverse results.

The chapter II various ideas concerning information are collected together from several sources. Applications in communication theory and statistical inference are considered.

In the final chapter, we are concerned with the information for discriminating between two Markov processes. After some preliminaries, a general result is proved for processes of the jump kind. We define a concept of infinitesimal information, and relate it to the discrimination information contained in an interval of time via an operator equation. The contents (except for some expository preliminaries) is original.

Two theorems, which could not be found in the literature in the desired form, but which digress from the main path of the dissertation, are included as appendices. The proofs are our own.

CHAPTER I

THE CONCEPT OF INFORMATION

1. Probability and Information.

Information is concerned with knowledge, and knowledge is derived from observations. In real life, all our observations are limited in precision, and our actions are finite in number. Mathematics has invented real numbers, which greatly simplify analysis, but no empirical scientist has ever observed "a real number". Rather he observes something in an interval whose size is determined by the precision of his instruments. But we can consider the real numbers as being "limits of precision", in a sense defined so as to "complete" the number system, either via Dedekind cuts, or Cauchy sequences.

Similarly, few Borel sets can be "measured". Rather we can only measure unions of a finite number of intervals. The "events" of classical probability theory are again limits (in some sense) of empirical events, invented for both aesthetic and utilitarian reasons. In the limiting process we give birth to the orphans called "sets of measure zero", which are both possible and impossible. But the fact is that these sets are not in fact observable, so it is a probabilistic enigma to call them events. Rather they are reminders of the fact that we put no upper limit on our possible precision. We can accomplish the same be talking about a measure algebra without atoms.

We will adopt the philosophy that any concept is the limit, in some sense, of that concept defined on finite systems. As much as possible

we will consider an algebra of events, rather than a "point set", as primitive. This "algebraic" approach to measure theory can be traced back to Carathéodory [4]. It has been applied to probability theory by Halmos [19], Birkhoff [2] and Kappos [25,26]. A fuller bibliography can be found in Kappos [26]. Information is defined on events and algebras of events, so that the "points" would for the most part be unnecessary baggage. However, there are times (particularly in the final chapter) where we would like to make use of standard results in analysis, and then we will represent our measure algebra in (Ω, A, μ) style.

1.1 There are several interpretations of the concept "probability" [3], but for the moment we will adopt the subjectivist one that probability is a measure of belief or expectation. Mathematically, we can consider a class of events, or observations; the "more likely" observations having a higher probability than the "less likely" ones. For any two events, we can conceive of their logical conjunction and logical disjunction; and if we adopt the convention of a sure event $\underset{\sim}{I}$ and an impossible event $\underset{\sim}{0}$, then our class has the structure of a Boolean algebra.

Let us call our class of events E. Our probability can be expressed as a function $p : E \to [0,1]$ such that $p(\underset{\sim}{0}) = 0$, $p(\underset{\sim}{I}) = 1$, $p\{E - \{\underset{\sim}{0},\underset{\sim}{I}\}\} \subset (0,1)$ and $A \wedge B = \underset{\sim}{0} \Rightarrow p(A \vee B) = p(A) + p(B)$, and hence the couple (E,p) defines a measure algebra. Denoting by "+" the symmetric difference on E, we find that $\rho(A,B) = p(A+B)$ is a metric, and to make things neater we usually talk about $\overline{E}$, the completion of E. p can be extended to $\overline{p}$ on $\overline{E}$ by continuity, and it follows that $\overline{p}$ is continuous in both the metric and the stronger, order topology [26, chapter 2].

Note that we have required the probability to be strictly positive, i.e., there is only one event of measure zero, the impossible event. (Such a convention is analogous to the situation in set theory, where there is only one "empty" set.) In relating this approach to the classical theory, it must be borne in mind that our events are modulo the σ - ideal of null sets - i.e., we identify sets which differ only by a set of measure zero.

A "probability" which is not strictly positive we will call improper. A "probability" for which $p(I) < 1$ we will call defective.

The completed algebra $\overline{E}$ is a σ - algebra, and the probability $\overline{p}$ is σ - additive. All probability algebras in this dissertation will be assumed to be complete in the above sense, and hence σ - algebras.

If the elements of $\overline{E}$ are thought of as propositions, then $\underset{\sim}{I}$ represents a tautology, $\underset{\sim}{0}$ a contradiction, and the other elements are contingent propositions. Elements which are "smaller" (in the lattice sense) represent more specific propositions, and people are often interested in making their statements more specific.

1.2 Probability is, of course, a monotone function on $\overline{E}$, and it is reasonable to expect that any function on $\overline{E}$ measuring information to be monotone decreasing, as one would think that more specific statements convey more information. We can draw a closer link between the concept of probability and information if we make a vague appeal to psychology, and maintain that the occurrence of an unexpected event (or equivalently, the confirmation of a dubious proposition) appears to convey more "information"

than the occurrence of an event we were expecting.

One can thus postulate that a measure of information should be a decreasing function of the probability. It is reasonable to expect that the information provided by independent events should be additive, hence a reasonable measure of the information of an event would be

$$J(A) = -\log p(A) \quad . \tag{1.2.1}$$

1.3 One can deduce (1.2.1) axiomatically [24]: Let J be a non-negative, decreasing function on $\overline{E}$, such that $J(\underset{\sim}{0}) = \infty$, $J(\underset{\sim}{I}) = 0$. Let us also postulate the existence of a finite or countable family of subalgebras $\{F_\alpha\}$ which is such that for any family of events $\{A_\alpha\}$ such that $A_\alpha \in F_\alpha$

$$\bigwedge_\alpha A_\alpha \neq \underset{\sim}{0} \quad . \tag{1.3.1}$$

Such a family is called M - independent. Intuitively, this means that elements of different F_α's are mutually compatible, so that presumably information about the one gives no information about the others.

Suppose now that J satisfies the following axioms:

For $$A_\alpha \in F_\alpha \ , \quad J(\bigwedge_\alpha A_\alpha) = \sum_\alpha J(A_\alpha) \quad . \tag{1.3.2}$$

There is an operation $\top$ on the positive extended real axis, such that for any $A,B \in \overline{E}$ such that $A \wedge B = 0$

$$J(A \vee B) = J(A) \top J(B) \quad . \tag{1.3.3}$$

If $s \neq r \neq t$, $A_\alpha \varepsilon \mathcal{F}_\alpha$ then

$$J(A_r) + [J(A_s \wedge A_t) \top J(A_s \wedge A_t^c) \top J(A_s^c \top A_b)$$

$$= [J(A_r) + J(A_s \wedge A_t)] \top [J(A_r)+J(A_s \wedge A_t^c)] \top [J(A_r)+J(A_s^c \wedge A_t)] \quad . \qquad (1.3.4)$$

Axiom (1.3.4) is somewhat technical and represents a limited distributive law.

Not all operations $\top$ make the axioms consistent. If we also require that we can assign information values within any sub-algebra $\mathcal{F}_\alpha$ with no regard to the values in the others, then $\top$ can only take on two forms.

$$x \top y = \inf(x,y) \qquad (1.3.5)$$

$$x \top y = -c \log (e^{-x/c} + e^{-y/c}) \quad . \qquad (1.3.6)$$

If $\top$ is of the form (1.3.5), then there exists no probability measure p on $\overline{\mathcal{E}}$ such that J can be represented as a strictly decreasing left-continuous function of p . If $\top$ is of the form (1.3.6), then there exists a probability measure p on $\overline{\mathcal{E}}$ such that

$$J = -c \log p \quad . \qquad (1.3.7)$$

1.4 Equation (1.2.1) defines the information gained by the confirmation of an uncertain event. We can generalize this formula if we say that after our experiment, the probability of Λ is now $q(A)$. (This generalization makes sense in either the subjectivist or logical interpretations of prob-

ability, but it has no frequentist interpretation.) Then we can say that the information transmitted by this change of probability is

$$J(A) = \log q(A) - \log p(A) \quad . \tag{1.4.1}$$

This quantity is properly described as information in favour of A, for if A^c is confirmed, and hence $q(A) = 0$, we have $J(A) = -\infty$; but we have not lost information about A, we have merely lost information in favour of A.

In (1.4.1) $q(A)$ is different from unity only if we do not directly observe A. We observe some other event B, and we know the stochastic link between A and B, and hence we can calculate $q(A)$; i.e., if we observe B, then $q(A)$ is the conditional probability $p(A|B) = p(A \wedge B)/p(B)$.

Formula (1.4.1) also has another interpretation (which can be reduced to the former under a Bayesean model). Note that it is in the form of a log likelihood ratio, and hence we can interpret it as the information in favour of q against p, provided by the observation A. The connection between these two interpretations is formal more than conceptual, but we will see a closer connection in the sequel. We will call the second interpretation discrimination information [28,29].

2. Sub-algebras and Conditioning.

So far we have been concerning ourselves with individual events. Let us now extend our horizons.

2.1 If $\mathcal{B}$ is a family of events in $\mathcal{A}$ such that $B_i \wedge B_j = \underset{\sim}{0}$ if $i \neq j$ and $\bigvee_{B \in \mathcal{B}} B = \mathcal{A}$ and such that $\underset{\sim}{0} \notin \mathcal{B}$, we will call $\mathcal{B}$ a <u>partition</u> in $\mathcal{A}$. It follows that $\mathcal{B}$ is at most countable. If $\mathcal{B}$ is finite we will speak about a finite partition.

Every partition generates a sub-algebra consisting of the join of its atoms, and conversely, every atomic sub-algebra defines a partition. We will often use the terms partition and atomic sub-algebra ambiguously. Of course, every finite sub-algebra is automatically atomic.

We can intuitively think of a partition as representing an experiment to determine which of the atomic propositions B in $\mathcal{B}$ is true.

2.2 We can define a (simple) random variable on a finite probability algebra as a function on the set of its atoms. We can also define elementary random variables as functions on the atoms of a countable partition. Note that the Boolean algebra generated by a countably infinite partition is, except in trivial cases, uncountable. We can define the usual function space operations (sum, product, scalar multiplication, positive, negative part) on the set of <u>all</u> elementary random variables, where if f and g are defined on $\mathcal{B}$ and $\mathcal{C}$ respectively, then $f \,\omega\, g$ is defined on $\mathcal{B} \vee \mathcal{C}$ for any operation ω.

With these definitions, the set $\underset{\sim}{E}(A)$ of all elementary random variables on $\mathcal{A}$ becomes a vector lattice, and we obtain the set of all random variables $\underset{\sim}{V}(A)$ as the order-completion of $\underset{\sim}{E}(A)$. It is well known that a non-negative measurable function (in the classical sense) is

the monotone limit of simple functions [37, p. 224], so that Kappos' approach is equivalent to the standard analytical approach.

Again, we can define expectation in the obvious way on $\underset{\sim}{E}(\mathcal{A})$ and extend it to $\underset{\sim}{L}(\mathcal{A}) \subset \underset{\sim}{V}(\mathcal{A})$ by continuity, following the Daniell approach [26, chapter V].

2.3 A Boolean algebra is, of course, a Boolean ring. If $\mathcal{B}$ is a subring of $\mathcal{A}$, it may be a Boolean algebra per se, although it is not a subalgebra of $\mathcal{A}$. This is true in particular if $\mathcal{B}$ is finite, or if $\mathcal{B}$ is a principal ideal. Principal ideals of a Boolean algebra are of the form $\mathcal{B} = \{A \in \mathcal{A} : A \leq B\}$ for some B. They will be denoted by $B\mathcal{A}$.

We will denote by $\underset{\sim}{F}(\mathcal{A})$ and $\underset{\sim}{R}(\mathcal{A})$ respectively the sets of all sub-σ-algebras and sub-σ-rings which are algebras. They are both complete lattices [2, p. 49] with greatest element $\mathcal{A}$, and least elements respectively $\mathcal{T} = \{\underset{\sim}{0},\underset{\sim}{I}\}$ and $\mathcal{N} = \{\underset{\sim}{0}\}$.

If $\underset{\sim}{C}$ is any class of Boolean σ-algebras which is a lattice under the operations

$$\mathcal{A} \wedge \mathcal{B} = \mathcal{A} \cap \mathcal{B}$$

$$\mathcal{A} \vee \mathcal{B} = \text{smallest } \sigma\text{-algebra containing } \mathcal{A} \text{ and } \mathcal{B}$$

we will call it a lattice of σ-algebras. It may not have a greatest, or a least element.

We will say that $\underset{\sim}{C}$ is a hereditary class if $\mathcal{A} \in \underset{\sim}{C} \Rightarrow \underset{\sim}{R}(\mathcal{A}) \subset \underset{\sim}{C}$.

We will require the following condition for any Boolean σ-algebra $\mathcal{A}$ we are studying:

> There exists an increasing sequence of finite sub-algebras $\mathcal{B}_n$ such that
>
> $$\bigvee_{n=1}^{\infty} \mathcal{B}_n = \mathcal{A} \quad . \tag{2.3.1}$$

Such algebras will be called separable. It follows that for any probability measure p on $\mathcal{A}$, $\mathcal{A}$ is also separable in the metric $\rho(A,B) = p(A+B)$ [20].

2.4 If $\underset{\sim}{0} \neq B \varepsilon \mathcal{B}$, then the principal ideal $B\mathcal{A} = \{A \varepsilon \mathcal{A} : A \leq B\}$ is also a Boolean algebra, and a probability p on $\mathcal{A}$ induces a probability p_B on $B\mathcal{A}$ defined by

$$p_B(C) = \frac{p(C)}{p(B)} \quad , \quad C \leq B \quad . \tag{2.4.1}$$

We can extend this measure as an (improper) probability on all of A by

$$p_B(A) = \frac{p(A \wedge B)}{p(B)} \tag{2.4.2}$$

the so-called conditional probability given B . We can make it a proper probability by considering $\mathcal{A}$ modulo $B^c\mathcal{A}$, which is isomorphic to $B\mathcal{A}$.

In this way, any finite partition $\mathcal{B}$ generates a family of probability algebras, one for each atom of the partition.

If f is any function defined on a hereditary class of probability algebras, then given an algebra $\mathcal{A}$ and a finite partition $\mathcal{B}$ in $\mathcal{A}$ we have a family of values

$$f[\mathcal{A}|B] = f(B\mathcal{A}) \tag{2.4.3}$$

indexed by the atoms B of $\mathcal{B}$. The above expression in fact defines a simple ($\mathcal{B}$-measurable) random variable, which we can denote by $f_{\mathcal{B}}[\mathcal{A}|\cdot]$, and we will denote the expectation by

$$f(\mathcal{A}|\mathcal{B}) = \sum p(B)\ f(B\mathcal{A}) \tag{2.4.4}$$

the summation extending over all atoms B of $\mathcal{B}$.

If $\mathcal{A}$ and $\mathcal{B}$ are sub-algebras of an algebra $\mathcal{C}$, but we don't have $\mathcal{B} \leq \mathcal{A}$, we can still define $f[\mathcal{A}|B]$ via (2.4.3) and hence $f(\mathcal{A}|\mathcal{B})$, where by $B\mathcal{A}$ we still mean all events of the form $B \wedge A$, $A \varepsilon \mathcal{A}$, although it is no longer an ideal of $\mathcal{A}$. It is easy to show that

$$f(\mathcal{A}|\mathcal{B}) = f(\mathcal{A} \vee \mathcal{B}|\mathcal{B}) \tag{2.4.5}$$

and that

$$f(\mathcal{A}|\mathcal{A}) = f(\mathcal{T}) \quad .$$

If $\mathcal{D}$ is any separable sub-algebra of $\mathcal{C}$, and if we have finite $\mathcal{B}_n \uparrow \mathcal{D}$, then we have a sequence of random variables

$$f_{\mathcal{B}_n}(\mathcal{A}|\cdot) \tag{2.4.6}$$

and if this sequence has a unique limit (in some sense) then we can talk of the random variable

$$f_{\mathcal{D}}(A|\cdot) \qquad (2.4.7)$$

and its expectation $f(A|\mathcal{D})$.

We will adopt the following terminology: The random variable will be called f conditioned by $\mathcal{D}$, its expectation will be called f conditional on $\mathcal{D}$.

2.5 Conditional probability with respect to a non-atomic sub-algebra can be defined as a Radon-Nikodym derivative.

For any A , $p_A(B) = p(A \wedge B)$ is a measure on $\mathcal{B}$, absolutely continuous with respect to p restricted to $\mathcal{B}$, hence the Radon-Nikodym derivative, $p(A|\mathcal{B})$ exists [26, p. 144].

Recall that we have only one impossible event, so that essentially we are identifying functions equal almost everywhere. Hence the derivative is unique.

Furthermore, if $A = A_1 \vee A_2$ $\underset{\sim}{0} = A_1 \wedge A_2$ then

$$\begin{aligned} \mathcal{E}(I_B(p(A_1|\mathcal{B}) + p(A_2|\mathcal{B}))) &= p(A_1 \wedge B) + p(A_2|B) \\ &= p((A_1 \vee A_2) \wedge B) \\ &= p(A \wedge B) \qquad \forall \qquad B \in \mathcal{B} . \end{aligned}$$

Hence, because of uniqueness,

$$p(A_1|\mathcal{B}) + p(A_2|\mathcal{B}) = p(A|\mathcal{B}) \quad . \tag{2.5.1}$$

Also, if $A_n \uparrow A$, then $p(A_n|\mathcal{B})$ is increasing, and by monotone convergence, for any $B \in \mathcal{B}$

$$\begin{aligned} \mathcal{E}(I_B \sup_n p(A_n|\mathcal{B})) &= \sup_n \mathcal{E}(I_B\ p(A_n|\mathcal{B})) \\ &= \sup_n p(A_n \wedge B) \\ &= p(A \wedge B) \quad . \end{aligned} \tag{2.5.2}$$

Hence, again because of uniqueness,

$$\sup_n p(A_n|\mathcal{B}) = p(A|\mathcal{B}) \quad . \tag{2.5.3}$$

We can thus consider conditional probability as a continuous vector-valued measure.

Conditional expectation could be defined as expectation with respect to conditional probability (see [21] for integration with respect to vector measures), or directly in terms of Radon-Nikodym derivatives [26, p. 228].

2.6 Two events A and B are said to be independent if $p(A \wedge B) = p(A)\ p(B)$. Two algebras $\mathcal{A}$ and $\mathcal{B}$ are independent if for all $A \in \mathcal{A}$, $B \in \mathcal{B}$, A and B are independent. It follows that $\mathcal{A} \wedge \mathcal{B} = \mathcal{T}$.

Two algebras $\mathcal{A}$ and $\mathcal{B}$ are conditionally independent given a third algebra $\mathcal{C}$, if for all $A \varepsilon \mathcal{A}$, $B \varepsilon \mathcal{B}$,

$$p(A \wedge B|\mathcal{C}) = p(A|\mathcal{C})\, p(B|\mathcal{C}) \quad . \tag{2.6.1}$$

If $\mathcal{A}$ and $\mathcal{B}$ are independent, and finite, then the p_B of (2.4.2) are identical to p for all B. It hence follows that for any function f of section 2.4.

$$f(\mathcal{A}|\mathcal{B}) = f(\mathcal{A}) \quad . \tag{2.6.2}$$

If $\mathcal{C}$ is atomic, and $\mathcal{A}$ and $\mathcal{B}$ are conditionally independent given $\mathcal{C}$ then

$$f(\mathcal{A}|\mathcal{B} \vee \mathcal{C}) = f(\mathcal{A}|\mathcal{C}) \tag{2.6.3}$$

for the atoms of $\mathcal{B} \vee \mathcal{C}$ are of the form $B \wedge C$ and $p_{B \wedge C} = p_C$.

3. Entropy and Discrimination Information.

3.1 Let us suppose that we are in a situation where we are independently receiving information from a large number of algebras isomorphic to $\overline{\mathcal{E}}$. Equivalently, we could imagine that after having an event in $\overline{\mathcal{E}}$ confirmed, the situation changes, and our uncertainty is restored. It is then reasonable to ask about a measure of average or expected information. If $\mathcal{F}$ is a sub-algebra generated by a finite partition, then the expected information in $\mathcal{F}$ is naturally defined as

$$H(\mathcal{F}) = - \sum p(A) \log p(A) \quad , \tag{3.1.1}$$

the summation extending over all atoms A of $\mathcal{F}$. This definition also makes sense in the case that $\mathcal{F}$ is generated by a countable partition, but in this case $H(\mathcal{F})$ may be infinite. The quantity H is usually called entropy.

Definition (3.1.1) can be derived axiomatically, either presupposing a probability on $\overline{E}$, or defining the probability in terms of H.

We can define entropy axiomatically in terms of a probability. Such axioms were first presented by Shannon [39]. Various versions and simplifications have been summarized by Aczel [1]. Since the entropy is a function of the probability distribution, the axioms are often expressed in terms of n-tuples of positive numbers summing to unity. We will here rephrase them in terms of Boolean algebras.

Let the entropy H be a function from the category of all finite probability algebras to the real numbers. For any finite algebra $\mathcal{A}$, and a sub-algebra $\mathcal{B}$, we can define the entropy of $\mathcal{A}$ conditional on $\mathcal{B}$ according to (2.4.4).

It is reasonable to want,

$$H(\mathcal{A}) = H(\mathcal{B}) + H(\mathcal{A}|\mathcal{B}) \quad \text{for each} \quad \mathcal{B} \leq \mathcal{A} \quad . \tag{3.1.2}$$

It turns out that formula (3.1.2) along with the requirement that the entropy of a diatomic algebra be a Lebesgue measurable function of the probability is sufficient to characterize (3.1.1) up to a scale constant [30].

In fact, it is sufficient to require that (3.1.2) be true only for sub-algebras B, one of whose atoms is equal to the join to two atoms of A, the other atoms being identical to those of A.

3.2 We can also characterize (3.1.1) using a limited definition of probabilities [23].

Let $\underset{\sim}{F}$ be the category of all finite Boolean algebras. $\underset{\sim}{F}$ is a hereditary lattice of algebras according to section 2.3.

For any $A \varepsilon \underset{\sim}{F}$ let $\underset{\sim}{S}(A)$ represent the set of all sub-rings of A. If ϕ is any homomorphism from A into B, it induces a map ϕ_* from $\underset{\sim}{S}(A)$ into $\underset{\sim}{S}(B)$. A sub-class $\underset{\sim}{H} \subset \underset{\sim}{F}$ is designated as a class containing homogeneous algebras. Within this sub-class, isomorphic algebras are identified.

For any $H \varepsilon \underset{\sim}{H}$, we define a probability by

$$p_H(A) = \frac{N(AH)}{N(H)} \tag{3.2.1}$$

where $N(A)$ represents the number of atoms of A. Let $\underset{\sim}{G} = \bigcup_{H \varepsilon \underset{\sim}{H}} S(H)$. We will first define H only on $\underset{\sim}{G}$.

We require the properties that

(a) If $H \varepsilon \underset{\sim}{H}$ and ψ is an automorphism of H then

$$H(\psi_*(B)) = H(B) \quad \text{for all} \quad B \varepsilon \underset{\sim}{S}(H) . \tag{3.2.2}$$

(b) H is isotone on $\underset{\sim}{H}$. i.e., if $G, H \in \underset{\sim}{H}$ and $G \leq H$ then

$$H(G) \leq H(H) \quad . \tag{3.2.3}$$

(c) If $H \in \underset{\sim}{H}$ and G is a sub-algebra of H (not necessarily in $\underset{\sim}{H}$) then

$$H(H) = H(G) + H(H|G) \quad . \tag{3.2.4}$$

H can be used to define a metric. If A and B are isomorphic, let

$$\delta(A,B) = \inf_{\phi} \sup_{C \in \underset{\sim}{S}(A)} |H(C) - H(\phi_*(C))| \tag{3.2.5}$$

where ϕ ranges over all isomorphisms from A to B. Now, if A,B are not isomorphic, let $\rho(A,B) = 1$, otherwise let

$$\rho(A,B) = \frac{\delta(A,B)}{1 + \delta(A,B)} \quad . \tag{3.2.6}$$

Then ρ is a metric on $\underset{\sim}{G}$, and under this metric, H is continuous. If we form the completion of $\underset{\sim}{G}$ with respect to ρ, and extend H by continuity, it can then be shown that we can define a probability p on each $A \in F$, and that

$$H(A) = -c \sum_{A \alpha A} p_A(A) \log p_A(A) \quad , \tag{3.2.7}$$

the summation extending over all atoms A of A. Furthermore, these probabilities are consistent, in the sense that if B is a sub-ring of A, then

$$p_{\mathcal{B}} = \frac{p_{\mathcal{A}} \mid \mathcal{B}}{p_{\mathcal{A}}(\underset{\sim}{I}_{\mathcal{B}})} \quad , \tag{3.2.8}$$

where $\underset{\sim}{I}_{\mathcal{B}}$ is the maximal element of $\mathcal{B}$ and by $p_{\mathcal{A}}|\mathcal{B}$ we mean $p_{\mathcal{A}}|\mathcal{B}$ restricted to $\mathcal{B}$.

3.3 We can also talk about average information in the generalized sense of (1.4.1).

We can consider two finite sub-algebras $\mathcal{A}$ and $\mathcal{B}$, the first consisting of events of interest, but not directly observable, the second consisting of observable events. We want to express the average information in $\mathcal{A}$, transmitted via $\mathcal{B}$.

If we observe $B \varepsilon \mathcal{B}$, the gain in information is:

$$J_B(A) = \log \frac{p(A|B)}{p(A)} \quad , \tag{3.3.1}$$

so that the average information in favour of A when A obtains is:

$$J^*(A) = \sum p(B|A) \log \frac{p(A|B)}{p(A)} \tag{3.3.2}$$

Summing over all atoms B of $\mathcal{B}$. The average information in $\mathcal{A}$ via $\mathcal{B}$ is hence:

$$R(\mathcal{A},\mathcal{B}) = \sum_A p(A) \sum_B p(B|A) \log \frac{p(A|B)}{p(A)} \tag{3.3.3}$$

summing over all atoms A of $\mathcal{A}$ and B of $\mathcal{B}$.

A typical application is communication theory, where $\mathcal{A}$ represents transmitted symbols, and $\mathcal{B}$ represents received symbols. We know the probability distribution μ of the transmitted symbols, and we know the noise characteristics ν of the channel. ν is expressed as conditional probabilities $\nu(B|A)$ of reception given transmission. We then have $p(A) = \mu(A)$,

$$p(B) = \sum \nu(B|A)\ \mu(A)$$

summing over all atoms A of $\mathcal{A}$, and $p(B|A)$ can be determined from Bayes' rule. Expression (3.3.3) can then be rearranged as:

$$\begin{aligned} R(\mathcal{A},\mathcal{B}) &= -\sum_{A} \mu(A) \log \mu(A) + \sum_{B} p(B) \sum_{A} p(A|B) \log p(A|B) \\ &= H(\mathcal{A}) - H(\mathcal{A}|\mathcal{B}) \end{aligned} \tag{3.3.4}$$

the summations extending over all atoms A of $\mathcal{A}$ and B of $\mathcal{B}$, and we see that it is the difference of two entropies. $H(\mathcal{A}|\mathcal{B})$ is called the "equivocation of the channel", and $R(\mathcal{A},\mathcal{B})$ is the rate of transmission [11].

R has also been used as a measure of the information provided by an experiment in a Bayesean framework [31].

3.4 Similarly we can talk about the average discrimination information, as

$$I(p,q;\mathcal{A}) = \sum p(A) \log \frac{p(A)}{q(A)} \tag{3.4.1}$$

the summation extending over all atoms A of $\mathcal{A}$. I can be thought of as the average information in $\mathcal{A}$ in favour of p against q, when the actual probability is p.

It is easy to show that "rate of transmission" (3.3.3), (3.3.4) can be expressed as a discrimination. In fact

$$R(\mathcal{A},\mathcal{B}) = I(p,p^*;\mathcal{A}\vee\mathcal{B}) \tag{3.4.2}$$

where p^* is the probability defined by taking $\mathcal{A}$ and $\mathcal{B}$ to be stochastically indpendent.

It may be possible that events possible under q, are impossible under p, so that q may be defective considered as a probability on $\mathcal{A}$. If events are impossible under q but possible under p, then total knowledge can be gained, and we say that $I[p,q] = \infty$.

When the probabilities p and q are understood, we will write simply $I(\mathcal{A})$. When the algebra is understood, and we wish to indicate the probabilities, we will use square brackets $I[p,q]$.

3.5 So far we have confined ourselves to information - theoretic concepts on finite algebras. Let us now attempt to extend them to the infinite case. We will restrict ourself to algebras $\mathcal{A}$ which are separable.

Let us first consider the concept of conditional entropy $H(\mathcal{A}|\mathcal{B})$, where for now we require that $\mathcal{A}$ and $\mathcal{B}$ are finite sub-algebras of a probabilitiy algebra $\mathcal{B}$. This concept includes unconditional entropy since $H(\mathcal{A}) = H(\mathcal{A}|\mathcal{T})$, $\mathcal{T}$ being the trivial sub-algebra.

We have already noticed that H is increasing in its first argument. We will now show that it is decreasing in its second.

First, observe that

$$g(t) = \begin{cases} t \log t & t > 0 \\ 0 & t = 0 \end{cases}$$

is a convex function and hence:

$$\sum p(B)\, g(p(A|B)) \geq g\left(\sum p(B)\, p(A|B)\right)$$

$$= g[p(A)] \quad , \tag{3.5.1}$$

the summations extending over all atoms B of $\mathcal{B}$. Therefore,

$$\begin{aligned} H(\mathcal{A}|\mathcal{B}) &= - \sum \sum p(B)\, g(p(A|B)) \\ &\leq - \sum g(p(A)) \\ &= H(\mathcal{A}) \quad , \end{aligned} \tag{3.5.2}$$

the summations extending over all atoms A of $\mathcal{A}$ and B of $\mathcal{B}$. Further, we note that if $\mathcal{C} \leq \mathcal{B}$ then

$$H(\mathcal{A}|\mathcal{C}) = \sum p(C)\, H(C\mathcal{A})$$

and

$$H(\mathcal{A}|\mathcal{B}) = \sum p(B)\ H(B\mathcal{A})$$

$$= \sum p(C)\ H(C\mathcal{A}|C\mathcal{B}) \quad , \qquad (3.5.3)$$

the summations extending over all atoms B of $\mathcal{B}$ and C of $\mathcal{C}$. Hence,

$$H(\mathcal{A}|\mathcal{C}) \geq H(\mathcal{A}|\mathcal{B}) \quad . \qquad (3.5.4)$$

We can thus define, for any algebra $\mathcal{B}$,

$$H(\mathcal{A}|\mathcal{B}) = \inf_{\mathcal{C}\leq\mathcal{B}} H(\mathcal{A}|\mathcal{C}) \quad , \text{ where } \mathcal{C} \text{ is finite.} \qquad (3.5.5)$$

Also, we can define for any algebra $\mathcal{A}$

$$H(\mathcal{A}|\mathcal{B}) = \sup_{\mathcal{C}\leq\mathcal{A}} H(\mathcal{C}|\mathcal{B}) \quad . \qquad (3.5.6)$$

3.6 We now turn our attention to discrimination information. It can readily be proved, again from the convexity of $t \log t$, that $I(\mathcal{A}|\mathcal{B})$ is increasing in its first argument.

Hence we can define, for any algebra $\mathcal{A}$,

$$I(\mathcal{A}|\mathcal{B}) = \sup_{\mathcal{C}\leq\mathcal{A}} I(\mathcal{C}|\mathcal{B}) \quad . \qquad (3.6.1)$$

However, I is not monotone in its second argument. It <u>is</u> true that if $\mathcal{A} \geq \mathcal{B} \geq \mathcal{C}$ that $I(\mathcal{A}|\mathcal{B}) \leq I(\mathcal{A}|\mathcal{C})$, but if $\mathcal{A}$ and $\mathcal{B}$ are incomparable it may be false. This is not surprising if one recalls that $I(\mathcal{A}|\mathcal{B}) = I(\mathcal{A}\vee\mathcal{B}|\mathcal{B})$.

A relatively simple counter-example can be found:

Let the atoms of $\mathcal{A}$ and $\mathcal{B}$ be respectively A_0 , A_1 and B_0 , B_1 , B_2 . Let p and q be defined by the following tables, on $\mathcal{A} \vee \mathcal{B}$.

p

	A_0	A_1
B_0	$\frac{1}{8}$	$\frac{1}{4}$
B_1	$\frac{1}{8}$	$\frac{1}{8}$
B_2	$\frac{1}{4}$	$\frac{1}{8}$

q

	A_0	A_1
B_0	$\frac{1}{6}$	$\frac{1}{6}$
B_1	$\frac{1}{6}$	$\frac{1}{6}$
B_2	$\frac{1}{6}$	$\frac{1}{6}$

Then we find that $I(\mathcal{A}) = I(\mathcal{A}|\mathcal{T}) = 0$ but

$$I(\mathcal{A}|\mathcal{B}) = \frac{5}{4} \log 2 - \frac{3}{4} \log 3$$

$$= .5425 \quad . \qquad (3.6.2)$$

It is quite in keeping with intuition that ancillary information from a previous experiment could improve the informativeness of a present experiment.

However, in spite of lack of monotonicity, if $\mathcal{B}_n$ is an increasing sequence of finite sub-algebras such that $\bigvee_{n=1}^{\infty} \mathcal{B}_n = \mathcal{B}$, then the sequences of random variables $I_{\mathcal{B}_n}(\mathcal{A}|\cdot)$ does converge, as does the

sequence of their expectations, hence we can unabmiguously talk about $I(\mathcal{A}|\mathcal{B})$ and $I_{\mathcal{B}}(\mathcal{A}|\cdot)$ whenever $\mathcal{B}$ is separable (see [18], proof of theorem 2.3).

The limit can in fact be represented as the discrimination information for the conditional probabilities of section 2.5.

4. Properties of Entropy and Discrimination Information.

We will here list some of the more significant aspects of these two quantities.

4.1 (a) Entropy is non-negative and isotone. It is zero only for the trivial algebra

$$\mathcal{A} \leq \mathcal{B} \Rightarrow H(\mathcal{A}) \leq H(\mathcal{B}) \tag{4.1.1}$$

$$H(\mathcal{A}) \geq 0 \quad H(\mathcal{A}) = 0 \quad \text{iff} \quad \mathcal{A} = \mathcal{T} \ . \tag{4.1.2}$$

(b) Entropy is conditionally additive

$$H(\mathcal{A}\vee\mathcal{B}) = H(\mathcal{B}) + H(\mathcal{A}|\mathcal{B}) \quad . \tag{4.1.3}$$

This can be derived for the finite case from (3.1.2), substituting $\mathcal{A}\vee\mathcal{B}$ for $\mathcal{A}$, and using (2.4.5). The property follows in the separable case by taking monotone limits. It follows from (2.6.2) that:

(c) Entropy is additive for independent subfields. $\mathcal{A}$ and $\mathcal{B}$ are independent

$$\Rightarrow H(\mathcal{A} \vee \mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B}) \quad . \tag{4.1.4}$$

(d) Conditional entropy is conditionally additive.

$$H(\mathcal{A} \vee \mathcal{B} | \mathcal{C}) = H(\mathcal{B} | \mathcal{C}) + H(\mathcal{A} | \mathcal{B} \vee \mathcal{C}) \quad . \tag{4.1.5}$$

For atomic algebras the result follows from (4.1.3) for each atom of $\mathcal{C}$. It follows in general by monotone limits.

(e) Both conditioned and conditional entropy are continuous in both arguments. If

$$\mathcal{B}_n \uparrow \mathcal{B} \quad ; \quad \mathcal{C}_n \uparrow \mathcal{C}$$

then

$$H_{\mathcal{B}_n}(A|\cdot) \to H_{\mathcal{B}}(A|\cdot), H_{\mathcal{C}_n}(A|\cdot) \to H_{\mathcal{C}}(A|\cdot)$$

$$H(\mathcal{A}|\mathcal{B}_n) \downarrow H(\mathcal{A}|\mathcal{B}), H(\mathcal{A}|\mathcal{C}_n) \uparrow H(\mathcal{A}|\mathcal{C})$$

$$H(\mathcal{B}_n|\mathcal{A}) \uparrow H(\mathcal{B}|\mathcal{A}), H(\mathcal{C}_n|\mathcal{A}) \downarrow H(\mathcal{C}|\mathcal{A}) \quad . \tag{4.1.4}$$

For proofs see [34] where our conditional entropy is called conditional information.

(f) Conditional entropy is anti-isotone in the conditioning algebra.

4.2 (a) Discrimination information is non-negative and isotone. It is equal to zero if and only if the two measure algebras are identical;

$$I(p,q;\mathcal{A}) \geq 0 \quad I(p,q,\mathcal{A}) = 0 \Leftrightarrow p(A) = q(A)$$

$$\forall \quad A \in \mathcal{A} \quad . \tag{4.2.1}$$

This follows from the convexity of $t \log t$, see [29].

(b) Conditional discrimination information is conditionally additive;

$$I(\mathcal{A} \vee \mathcal{B}) = I(\mathcal{B}) + I(\mathcal{A}|\mathcal{B}) \quad . \tag{4.2.2}$$

This has been proved (for separable $\mathcal{B}$) by Ghurye [18]. If $\mathcal{A}$ and $\mathcal{B}$ are independent then

$$I(\mathcal{A} \vee \mathcal{B}) = I(\mathcal{A}) + I(\mathcal{B}) \quad . \tag{4.2.3}$$

(c) Conditional discrimination information is discrimination information. i.e., $\exists \ q^*$ such that

$$I(p,q;\mathcal{A}|\mathcal{B}) = I(p,q^*;\mathcal{A}) \tag{4.2.4}$$

see [7], equation 3.3.

(d) If f is the Radon-Nikodym derivative of p with respect to q , then

$$I(p,q;\mathcal{A}) = E_p(\log f) \quad . \tag{4.2.5}$$

This expression is usually taken as a definition. This result has been demonstrated by Kolmogorov, Gelfand and Yaglom [27] and Ghurye [18].

(e) Conditional information characterizes sufficiently, i.e., $\mathcal{B} \leq \mathcal{A}$ $I(\mathcal{A}|\mathcal{B}) = 0$ iff $\mathcal{B}$ is a sufficient sub-field for the pair $\{p,q\}$.

(f) Discrimination information is continuous.

If $\mathcal{A}_n \uparrow \mathcal{A}$ then $I(\mathcal{A}_n) \uparrow I(\mathcal{A})$. (4.2.6)

Proof follows from the convexity of $t \log t$ and appendix 1.

4.3 Discrimination information, in measuring the ease in differentiating one probability distribution from another, in a sense measures a "distance" between probability measures. It is however, not a metric on the space of probability measures, for it is neither symmetric, not does it satisfy the triangle inequality.

However, one can define convergence of probability in terms of discrimination information, and this convergence is stronger then convergence in the total variation norm [8]. In general, however, the convergence structure is only a Frechet-V-space [16,17], and not even a topological space.

In classical statistical problems, our set of possible probabilities is parametrized by a convex set in Euclidean space, and thus has not only a metric, but also a differentiable structure. If we let θ_o represent the parameter for the null hypothesis, then $I[\theta_o,\theta]$ is a measure of

how easy it is to reject the null hypothesis under the alternative θ . The local behaviour of I about θ_o suggests how "secure" our statistical inferences regarding θ_o are. Since $I[\theta_o,\theta]$ is increasing away from θ_o , it is easy to see that the total differential, if it exists, must be zero. Hence the curvature would give an idea as to how quickly I moves away from its tangent plane, i.e., how quickly does the discrimination information increase. Let $D(\theta_o)$ be the matrix of second derivatives of $I[\theta_o,\theta]$ at θ_o . Then, provided that we assume sufficient regularity conditions, it is easily seen that $D(\theta_o)$ is Fisher's information matrix [29].

CHAPTER II

A BRIEF HISTORY

1. Information and Communication.

A concept central to communication theory is a communication channel, consisting of a set of inputs X, a set of outputs Y, and a line between them ν. For the moment we will leave these three components further unspecified.

1.1 Our first problem is to define what we mean by the amount of information which can be transmitted, that is, the information which can be conveyed by X. If X contains n symbols, then we can construct n^k sequences (or "super-symbols") from k copies of X, and as we would intuitively expect that k copies of X contain k times as much information, it seems reasonable to use $c \log n$ as a measure of the information potential of X.

However, let us suppose we have two channels, and in both cases X_i contains 2 symbols, 0 and 1. A message is sent along each channel once a second. However, in the first channel, 0 and 1 are sent with about equal regularity, but in the second one, most messages are 1, and a 0 occurs on the average only once a year. We would intuitively think that the second channel would be transmitting far less information.

It hence appears that the probabilities of the symbols being transmitted affect the amount of information, so that X would be completely

described as a probability space $(X,\mathcal{X},p)$.

Claude Shannon in 1948 [39] gave axioms which an information measure should satisfy, and derived from them the formula for entropy (I, 3.1.1). Shannon's axioms were stronger than the ones we gave in (I, 3.1). He assumed, that H , as a function of the individual probabilities for fixed n (the number of symbols) was continuous; and was increasing in n , if the probabilities were uniform; and essentially (I, 3.1.2). It is customary to use logarithms to base 2.

Entropy is maximum when the probabilities are uniform, and any "averaging process" (essentially a convolution) tends to increase it. These properties are what one would expect if entropy is thought of as a measure of disorder. The concept had been used in that context earlier in statistical mechanics [40].

If we consider the product set X^n with probability P^n , as the set of n-symbol sequences sent independently, then we have:

$$\forall \epsilon > r , \forall \delta > 0 \ \exists \ n \ni \forall n > m \ \exists \ C \subset X^n \ni$$

$$P^n(C^c) < \epsilon \quad \text{and} \quad \forall s \in C$$

$$|\log_2 P^n\{s\} + nH| < \delta \tag{1.1.1}$$

i.e., 2^{-nH} is the approximate probability, for large n , of each "reasonably probable" sequence of n elements.

1.2 If there is a one-to-one correspondence between the input symbols and output symbols, then we speak of a noise-less channel. In this case, no information is lost in transmission. The more usual case is where the link between the channels is stochastic. For each $x \in X$, there is a probability distribution $\nu(x,\cdot)$ on Y. The situation is now identical with (I, 3.3), and we can define $R(X,Y)$ as the rate of transmission. Note that the noise characteristic ν is assumed known.

The supremum of $R(X,Y)$ over all probabilities p on X is called the capacity of the channel. Coding theory is concerned with choosing symbols so as to maximize capacity.

1.3 Many real-life communication channels transmit signals which are continuous varying voltages, and hence are not expressible as one of a finite number of symbols. It would be useful to have an information measure for such channels as well. Unfortunately, the obvious quantity (3.5.6) is always infinite in this case, and hence is not very useful.

Shannon and Wiener [40] independently proposed the measure

$$- E(\log f) \tag{1.3.1}$$

(though with opposite signs,) f being the usual density of the probability distribution.

Discrimination information has been generalized by Ghurye [18] where p is any finite measure, absolutely continuous with respect to q, which is σ - finite, and (I, 4.2.5) holds. Hence, the entropy of a contin-

uous distribution p can be written as $I[p,q]$, q being Lebesque measure, in a sense the "most random" distribution.

Continuous entropy may be infinite, so it has no "maximum". However for all distributions concentrated on a bounded set, entropy is maximum if the distribution is uniform; for all distributions concentrated on the positive real axis with expectation λ it is maximum if the distribution is exponential λ ; and for all distributions with variance σ^2 , it is maximum if the distribution is normal σ^2 . Entropy is location invariant but it depends on the scale. It may be negative. Like discrete entropy, continuous entropy is increased by an "averaging" process.

1.4 The rate of transmission can be defined in the continuous case even when the entropy is infinite, as a discrimination information. cf. (I, 3.4.2). Let f be the density of the input signal, let

$$p(x,y) = f(x)\ \nu_x(y)$$

(ν being the noise characteristic, see section 1.2)

$$\mu(y) = \int f(x)\ \nu_x(y)\ dy$$

and

$$q = f\ \mu \quad ,$$

then

$$R(X,Y) = I[p,q] \quad . \tag{1.4.1}$$

2. Information and Inference.

2.1 The concept of information in the context of statistics was introduced by R.A. Fisher [13,14]. He defined a "sufficient statistic" as one which contains all the information in an experiment. This intuitive idea is tersely expressed by (I, 4.2e).

The concept was also used in the theory of estimation, as corresponding to reciprocal variance (of an estimator). The more dispersed is our estimator, the less efficient or informative it is.

If $\hat{\theta}$ is a maximum likelihood estimator, it is the solution of

$$\frac{\partial L(\theta)}{\partial \theta} = 0$$

where

$$L = \sum_{i=1}^{n} \log f_{\theta}(x_i) \tag{2.1.1}$$

f_{θ} being the density of each sample value x_i. If $\hat{\theta}$ is unbiased, and normally distributed, then we have

$$\left.\frac{\partial^2 L}{\partial \theta^2}\right|_{\theta=\hat{\theta}} = -\frac{1}{\sigma^2_{\hat{\theta}}} \tag{2.1.2}$$

so that we can use $-\frac{\partial^2 L}{\partial \theta^2}$ as a measure of the informativeness of $\hat{\theta}$.

We can also talk about expected information:

$$I(\theta) = E_{\theta}\,[-\frac{\partial^2 L}{\partial \theta^2}] = n \int \frac{1}{f}\,(\frac{\partial f}{\partial \theta})^2 = n\; i(\theta) \quad . \tag{2.1.3}$$

Fisher calls the quantity $i(\theta)$ the intrinsic accuracy of the density f_θ , and says it represents the maximum amount of information provided by a single observation [14, p. 709].

Under appropriate regularity conditions, we have the well-known Rao-Cramer inequality [35, 5 p. 477ff] which gives for an unbiased estimator:

$$\mathrm{var}(\hat{\theta}) \geq \frac{1}{n\, i(\theta)} \quad . \tag{2.1.4}$$

Equality holds if and only if the estimator is sufficient and normally distributed, so that Fisher's idea of maximum information is well founded. (2.1.4) can be extended to the multi-parameter case [35]. Letting I be the matrix

$$- E_\theta \left(\frac{\partial^2 \log f_\theta}{\partial\theta_i \; \partial\theta_j}\right)_{\theta=\hat{\theta}} \tag{2.1.5}$$

then we have that

$$\mathcal{D}(\hat{\theta}) \geq n^{-1} \; I^{-1} \tag{2.1.6}$$

$\mathcal{D}$ representing the dispersion matrix. We have already seen the relation between the matrix I and discrimination information (I, 4.3).

2.2 If $\hat{\theta}$ is a sufficient estimator, then we can write $f_\theta(x) = g_\theta(t) \cdot h(x|t)$, where g is the density of $\hat{\theta}$. Then,

$$\frac{\partial^2 \log f_\theta(x)}{\partial\theta^2} = \frac{\partial^2 \log g_\theta(x)}{\partial\theta^2} \quad . \tag{2.2.1}$$

Bearing this in mind, we can define, for any random vector T such that its one-parameter density is twice-differentiable, the information quantity:

$$I_\theta(T) = {}_\theta\left(\frac{\partial^2 \log g_\theta}{\partial\theta^2}\right) \quad . \qquad (2.2.2)$$

This quantity has certain attractive properties, [34, p. 5].

(a) $I_\theta(T) \geq 0$ equality holding iff T has the same distribution for all θ .

(b) $I_\theta(T) \leq I_\theta(X)$ if T is a statistic from X , equality holding only in the case of sufficiency.

(c) $I_\theta(T_1, T_2) = I_\theta(T_1) + I_\theta(T_2)$ if T_1 and T_2 are independent.

2.3 Fisher drew a parallel between his information and entropy [15, p. 47]. Thermodynamic processes which are irreversible are accompanied by an increase in entropy. Statistical processes which are irreversible (i.e., data processing) cannot increase information.

2.4 A complete theory of inference based on discrimination information has been developed by Kullback [29]. In it, the discrimination information is used as a pseudo-distance. Central to his development is an exponential family of distributions "closest" to a "null" distribution of all distributions yielding the same expected value for a statistic.

2.5 The calculation of a statistic can be considered as a communications channel as in section 1 which is ambiguous but otherwise noiseless - i.e., the measures $\nu(x,\cdot)$ are degenerate. Csiszar [7] introduced the idea of a noisy channel in statistics, which may be realized in the case where an observation has an error in addition to the intrinsic error of the experimental material. He calls it the case of indirect observation.

Not surprisingly, an indirect observation can only decrease the discrimination information. If the decrease is zero, the channel can be called sufficient. If the decrease is less than ε , the channel can be called ε - sufficient. (However, if the information in both the direct and indirect observations is infinite, the "decrease" is undefined.) The same applies a fortiori to statistics and subfields. If a sufficient channel does not exist, or is expensive to construct, and ε - sufficient one may be adequate.

3. Various Tangents.

3.1 Renyi [36] has considered generalizations of entropy and discrimination information by relaxation of the axioms.

By replacing the conditional additivity (I, 3.1.2) by independent additivity (I, 4.1.4) he deduces that the quantities

$$H_\beta(\mathcal{A}) = \frac{1}{1-\beta} \log \left(\sum [P(A)]^\beta \right)$$

$$1 \neq \beta > 0 \quad , \tag{3.1.1}$$

the summation extending over all atoms A of $\mathcal{A}$.

also satisfy the axioms. It is easily shown that

$$\lim_{\beta \to 1} H_\beta(\mathcal{A}) = H(\mathcal{A}) \quad . \tag{3.1.2}$$

Renyi's treatment includes defective probabilities as well. In this case the quantity in (3.1.1) is divided by $P(\underset{\sim}{I})$.

Similarly, Renyi generalized discrimination information to

$$I_\beta(\mathcal{A}) = \frac{1}{\beta-1} \log \sum_{A \varepsilon \mathcal{A}} q(A) \left[\frac{p(A)}{q(A)}\right]^\beta \quad , \quad \beta \neq 1 \quad , \tag{3.1.3}$$

the summation extending over all atoms A of $\mathcal{A}$. This quantity again satisfies the independent additivity property (I, 4.2.3).

3.2 Discrimination information was seen, in the general case, to be the supremum of discrimination information over all finite sub-algebras.

An analogous quantity was defined by Ghurye [18] for any finite measure p dominated by a σ-finite measure q defined on $\mathcal{A}$, and any convex function f defined on $[0,\infty]$, as

$$I_f(\mathcal{A}) = \sup_{\mathcal{B} \leq \mathcal{A}} \sum q(B) \, f\left(\frac{p(B)}{q(B)}\right) \tag{3.2.1}$$

where the summation extends over all atoms B of the atomic subalgebras $\mathcal{B}$ of $\mathcal{A}$.

If $\phi = \frac{dp}{dq}$ then we have the analogue of (I, 4.2.5)

$$I_f(\mathcal{A}) = \int f \circ \phi \; dq \quad . \tag{3.2.2}$$

We have seen that continuous entropy is an example of this generalization. We will need it again in chapter III.

3.3 Two of the most significant properties of discrimination information are that it is monotonically increasing: ($B \leq A \Rightarrow I(B) \leq I(A)$; but is strictly increasing only for non-sufficient sub-algebras, i.e., $I(B) = I(A)$ if B is a sufficient sub-algebra). It turns out that convexity of $t \log t$ is the only assumption needed to prove these two facts. Csiszar [7] has hence defined for any convex function f, the f-divergence, between the probability measures p and q, defined by (3.2.2). To take care of the situation where p is not dominated by q, he defined,

$$0 \cdot f\left(\frac{0}{0}\right) = 0 \quad \text{and} \quad 0 \cdot f\left(\frac{a}{0}\right) = \lim_{\epsilon \to 0^+} \epsilon f\left(\frac{a}{\epsilon}\right) = a \lim_{u \to \infty} \frac{f(u)}{u} . \tag{3.3.1}$$

There may be situations where some f-divergences are more meaningful than discrimination information. It has already been mentioned that the neighbourhood systems defined by I on the space of measures need not define a topological space. However, if both $f(0)$ and $\lim_{t \to \infty} \frac{f(t)}{t}$ are finite and f is strictly convex at 1, then I_f does define a topology, equivalent to the variation difference, the latter itself being an f-divergence with $f(t) = |t-1|$ [8].

If p is not dominated by q, then discrimination information is infinite. However, if an event is impossible under q, yet has a positive but very small probability under p, we may not want to think of this "discrimination distance" as so large, as in fact the two distributions may be very difficult to tell apart by experiment. By appropriate choice of f, we can allow I_f to remain finite in such a case.

CHAPTER III

INFORMATION IN MARKOV PROCESSES

1. Basics of Markov Theory.

1.1 Let us now suppose that we have a family of sub-algebras of a probability algebra $\mathcal{A}$, which has a temporal structure. Let T be a totally ordered set (which is usually assumed to be either the non-negative integers, or the non-negative real line). For every $t \in T$, we will suppose we have a sub-algebra $\mathcal{B}_t$, called the algebra of events observed at epoch t. Also, for any closed interval $[s,t]$, we have an algebra of events $\mathcal{C}_{s,t}$, called the algebra of events observed between epochs s and t. As we will be dealing solely with separable algebras, we will assume that each $\mathcal{B}_t$ and each $\mathcal{C}_{s,t}$ is separable. We will, in fact, assume a stronger separability condition:

If D is a dense countable subset of (s,t) then

$$\mathcal{C}_{s,t} = \bigvee_{d \in D} \mathcal{B}_d \quad . \tag{1.1.1}$$

This latter condition, of course, is trivial in the case that T is countable.

If we have a continuous probability p defined on $\mathcal{A}$, we will call the system $(\mathcal{A},T,\mathcal{B}_t,\mathcal{C}_{s,t},p)$ a separable stochastic process. We will say that it has the Markov property if for any t and $s > t$, we have:

The algebras $\mathcal{C}_{o,t}$ and $\mathcal{C}_{t,s}$ are conditionally independent

given $\mathcal{B}_t$.

We will say that is is temporally homogeneous if, for any s and t and h $(s < t)$

$$p(E|\mathcal{B}_s) = p(F|\mathcal{B}_{s+h}) \quad E \in \mathcal{B}_t \quad F \in \mathcal{B}_{t+h} \quad . \tag{1.1.2}$$

We will concern ourselves exclusively with temporally homogeneous processes having the Markov property.

The quantities in 1.1.2 are called transition probabilities, and our separability condition 1.1.1 ensures that any event in any $\mathcal{C}_{s,t}$ has a unique probability defined in terms of them, conditional on $\mathcal{B}_o$.

1.2 The preceding was a brief introduction to stochastic processes in the language of Chapter I. In order to make use of other results, without the need for lengthy reformulation, we will return to more standard notation. We will assume that the algebras $\mathcal{B}_t$ are all isomorphic, and hence can be represented by a family of random variables with values in the same space.

Let $(\Omega,\mathcal{A},p)$ be our basic probability space, and let $(E,\mathcal{E})$ be another measurable space which we call the state space. For each $t \in T$ let X_t be a measurable function from Ω to E . Our algebras $\mathcal{B}_t$ then are simply $X_t^{-1}(\mathcal{E})$, and the values which X_t takes on are our observables.

Two processes X_t and Y_t on the same state space are said to be equivalent if the finite-dimensional distributions are the same. i.e. if $\forall\, n$, $\forall\, t_1,\cdots,t_n \in T\ \forall\, F_1,\cdots,F_n \in \mathcal{E}$

$$p_1\{X_{t_j} \in F_j \ ;\ j = 1,n\} = p_2\{Y_{t_j} \in F_j \ ;\ j = 1,n\} \ . \qquad (1.2.1)$$

We will be concerned exclusively with finite-dimensional distributions and their limits, so that we will essentially identify equivalent processes. As every stochastic process is equivalent to a separable process [32, IV, T29], our separability condition is hence no restriction. (Although our definition of separability appears more restrictive than Meyer's, I conjecture they are equivalent if one assumes continuity in probability. See [9, II, Theorem 2.2].)

1.3 Transition probabilities are usually expressediin terms of Markovian kernels [12, 10, 33]

$$P_t(x,F) = p(X_{s+t} \in F \mid X_s = x)$$

$$= p(X_{s+t}^{-1}(F) \mid \mathcal{B}_s) \ . \qquad (1.3.1)$$

Given an initial distribution p_o , we can define the distribution at any time t by

$$p(X_t \varepsilon F) = \int p_o(dx)\ P_t(x,F) \qquad (1.3.2)$$

hence we can see that the Markov kernels act as operators on the linear

space of measures on E, which leave the subset of probability measures invariant.

The also define operators on the dual space of functions on E, defined by

$$(P_t^* f)(x) = \int P_t(x,dy)\ d(y) \quad .$$

1.4 The operators $\{P_t^*\}$ form a semi-group of transition operators. That is, they satisfy the Chapman-Kolmogorov relation

$$P_t^* P_s^* = P_{t+s}^* \quad , \tag{1.4.1}$$

they take positive functions into positive functions, and they leave constant functions invariant. From (1.4.1) we can see that they commute with each other. They are of norm 1.

The operators can be embedded in a Banach algebra, in which analogues of classical algebraic and analytic procedures can be developed [21] including limits, differentiation and integration. There are two limits that concern us. We say that the operator sequence $\{T_n\}$ converges to T uniformly or strongly as, respectively

$$||T_n - T|| \to 0 \tag{1.4.2}$$

$$||T_n f - Tf|| \to 0 \quad \forall\ f \in D \tag{1.4.3}$$

D being the domain of the operators. Uniform convergence implies strong

convergence.

Semi-groups of transition operators such that $P_o^* = I$ and that $P_t^* \to I$ strongly as $t \downarrow 0$ are called the Feller semi-groups [33], and are in fact (strongly) continuous everywhere. We will deal exclusively with such semi-groups, and in fact we will suppose that the transition probabilities are conservative, which is to say they are non-defective probability measures.

If the limit

$$A^* = \lim_{t \downarrow 0} \frac{P_t^* - I}{t} \tag{1.4.4}$$

exists uniformly, then we say that the semi-group is (uniformly) differentiable, and from this fact follows the Kolomogrov background equation:

$$\frac{dP_t^*}{dt} = A^* P_t^* \quad . \tag{1.4.5}$$

In fact, the semi-group (P_t^*) can be represented as:

$$P_t^* = \sum_{k=0}^{\infty} \frac{(A^* t)^k}{k!} = e^{A^* t} \quad . \tag{1.4.6}$$

A^* is called the infinitesimal generator of the semi-group.

If the limit in (1.4.4) does not exist uniformly, but if it exists strongly in some subspace C dense in D (i.e.

$$A^* f = \lim_{t \downarrow 0} \frac{P_t^* f - I}{t} \quad \forall\ f \in C \quad , \tag{1.4.7}$$

the limit being in the norm), we still say that A^* is the infinitesimal generator. (1.4.5) holds still [although $\frac{dP_t{}^*}{dt}$ is now only a strong derivative] but (1.4.6) need not, since A^* may be an unbounded operator, so that the exponential function is not defined. However, the semi-group can be approximated by semi-groups of the form (1.4.6) [12]. We also have the inverse of (1.4.5), viz.

$$P_t{}^* - I = A^* \int_0^t P_s{}^* \, ds \quad . \tag{1.4.8}$$

This relation is true on D if (1.4.4) exists uniformly, and is still true on C if only (1.4.7) holds [10,21].

1.5 An important class of Markov processes is that of the "step" kind. That is, the process remains constant for a random length of time, and then jumps to another state, where it again remains for a random length of time. Only a finite number of jumps occur in any finite interval.

There are three limit expressions which figure in the study of processes of this kind.

$$\lim_{t \downarrow 0} P_t(x,\{x\}) = 1 \tag{1.5.1}$$

$$\lim_{t \downarrow 0} \frac{P_t(x,\{x\}) - 1}{t} = A_o(x,\ x\ ) \tag{1.5.2}$$

$$\lim_{t \downarrow 0} \frac{P_t(x,F)}{t} = A_1(x,F) \quad F \varepsilon \{x\}^c \ \mathcal{E} \quad . \tag{1.5.3}$$

Convergence in any of the three can be pointwise or uniformly in x, and in (1.5.3) the convergence can be pointwise or uniform in F.

(1.5.1) expresses the fact that if $X_s = x$, then for t sufficiently small $X_{s+t} = x$ (almost surely). It is a necessary condition for the sample functions to be step functions.

Uniform convergence of (1.5.1) is equivalent to the boundedness of A_o and implies uniform convergence of (1.5.2) and uniform convergence of (1.5.3) in F. Furthermore, A_1 is in fact a measure, so that $A = A_0 + A_1$ can be considered a signed measure, for each x, with $A(x,E) = 0$; and $A_1(x,F)$ is a measurable function for each F [9, VI. 2].

If in fact the convergence in (1.5.3) is also uniform in x, then it follows that (1.4.4) exists uniformly, and

$$A^* f(x) = \int A(x,dy)\ f(y) \tag{1.5.4}$$

(see appendix 2).

Even if (1.4.4) does not hold uniformly, the operator A^* will be bounded provided that the function A_o is.

2. Two Simple Examples.

We will now consider discrimination information in a temporally homogenous Markov process.

Let us suppose that we have a separable stochastic process $(\mathcal{A},T,\mathcal{B}_t,\mathcal{C}_{s,t},p)$. If q is another probability on $\mathcal{A}$, we can consider $I[p,q]$, the discrimination information in favour of p against q contained in various sub-algebras of $\mathcal{A}$. In particular we will be interested in the information contained in the sub-algebras $\mathcal{B}_t$ and $\mathcal{C}_{o,t}$.

Csiszar has considered discrimination information between two Markov chains with different initial probabilities, and the same transition probabilities [6]. Not surprisingly, $I(\mathcal{B}_t)$ decreases with t , in fact, if the chain is recurrent, irreducible and aperiodic, it converges to zero, hence proving ergodicity.

We will take the opposite approach. We will assume that the initial subfield $\mathcal{B}_o$ is given, and will be concerned with the discrimination information between two families of transition probabilities.

First, we will calculate the discrimination information directly, for two simple cases.

2.1 Suppose we have two Poisson processes starting with $X_o = 0$, with probabilities p_{λ_1} and p_{λ_2} , and with intensity parameters λ_1 and λ_2 respectively. We want to find the discrimination information in favour of p_{λ_1} against p_{λ_2} (i.e. in favour of the hypothesis $\lambda = \lambda_1$ against the hypothesis $\lambda = \lambda_2$) in the subfield $\mathcal{C}_{o,t}$.

For any integer n , let $S_{kj}^{(n)}$ represent the measurable set

$$\{\omega : X(\tfrac{k}{n} t,\omega) - X(\tfrac{k-1}{n} t,\omega) = j\} \tag{2.1.1}$$

and let $J^{(n)}$ be the σ-algebra generated by these sets. The atoms of this algebra are

$$\bigcap_{k=1}^{n} S_{kj_k}^{(n)} = C_{\underset{\sim}{\eta}}^{(n)} \tag{2.1.2}$$

where $(j_1,\cdots,j_n) = \underset{\sim}{\eta} \in N_o^n$.

Now, because the process has independent increments, we have

$$p_{\lambda_r}(C_{\underset{\sim}{\eta}}^{(n)}) = \prod_{i=1}^{n} \exp\left(-\frac{\lambda_r t}{n}\right)\left(\frac{\lambda_r t}{n}\right)^{j_i} \quad , \quad r = 1,2 \tag{2.1.3}$$

Hence,

$$I(p_{\lambda_1},p_{\lambda_2};J^{(n)}) = \sum_{C_{\underset{\sim}{\eta}}^{(n)}} p_{\lambda_1}(C_{\underset{\sim}{\eta}}^{(n)}) \log\left[\frac{p_{\lambda_1}(C_{\underset{\sim}{\eta}}^{(n)})}{p_{\lambda_2}(C_{\underset{\sim}{\eta}}^{(n)})}\right] \tag{2.1.4}$$

$$= (\lambda_2-\lambda_1)t + \lambda_1 t \log\frac{\lambda_1}{\lambda_2} \quad .$$

Note that n does not figure in (2.1.4), so that using (1.1.1) and (I,4.2.6) we obtain

$$I(p_{\lambda_1},p_{\lambda_2};C_{o,t}) = (\lambda_2-\lambda_1)t + \lambda_1 t \log\frac{\lambda_1}{\lambda_2} \quad . \tag{2.1.5}$$

2.2 Next, let us suppose that we have two Brownian motions starting at the origin with differing diffusion rates a and b . Again we want the discrimination information in favour of the hypothesis $\sigma_t^2 = at$ against the hypothesis $\sigma_t^2 = bt$, contained in $C_{o,t}$.

Let $\mathcal{J}^{(n)}$ represent the algebra

$$\bigvee_{k=1}^{n} \mathcal{B}_{\frac{kt}{n}} \quad .$$

The discrimination information between two central normal distributions with different variances can readily be calculated as

$$I[\sigma_1^2,\sigma_2^2] = \frac{1}{2}\left(\log \frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - 1\right) \tag{2.2.1}$$

hence we see that the information depends only on the ratio of the variances.

Thus, the information in the sub-field $\mathcal{B}_t$ is equal to

$$\frac{1}{2}\left(\log \frac{a}{b} + \frac{a}{b} - 1\right) \tag{2.2.2}$$

and in $\mathcal{J}^{(n)}$ it is equal to (using the property of independent increments)

$$\frac{n}{2}\left(\log \frac{a}{b} + \frac{a}{b} - 1\right) \quad . \tag{2.2.3}$$

Letting $n \to \infty$, we see that the information in $\mathcal{C}_t$ is infinite.

3. A General Formulation.

3.1 Let us first define some notation.

For any t let $\phi(t,x)$ be $I_{\mathcal{B}_o}(\mathcal{B}_t|x)$, the discrimination information in $\mathcal{B}_t$ conditioned by $\mathcal{B}_o$. It is the discrimination information between the transition probabilities with lag t, and is for given t

a random variable on $\mathcal{B}_o$, i.e. a function on E .

By $\psi(t,x)$ we will denote $I_{\mathcal{B}_o}(\mathcal{C}_{o,t}|x)$ the information in $\mathcal{C}_{o,t}$ conditioned by $\mathcal{B}_o$. It is also a function on E . If we know the initial distribution π , then

$$I(\mathcal{C}_{o,t}) = I(\mathcal{B}_o) + \int \psi(t,x)\, d\pi(x) \quad . \tag{3.1.1}$$

Obviously ψ is increasing in its first argument, and

$$\phi \leq \psi \quad . \tag{3.1.2}$$

An important case is that in which $\psi < \infty$ and equality holds, for then the observation at epoch t is as informative in discriminating the two probabilities as all events occuring beforehand.

We noticed this case for the Poisson process.

We will require the following regularity condition on ϕ : For every t , $\phi(t,\cdot) \varepsilon D$, the domain of the transition operators.

In order to calculate ϕ directly, we must have explicit expressions for the two sets of transition probabilities P_t and Q_t . For many concrete processes of the step kind, the parameters are defined in terms of the intensities defined in 1.5; and expressions for the transition probabilities can be very cumbersome. We will derive an operator equation for ψ , which does not involve ϕ .

We will maintain the following notation: P_t will be the transition probability of the process under study, its associated operator will be P_t^*, A^* the infinitesimal generator. If the intensities of 1.5 exist, they will be denoted by A_o and A_1, $A = A_o + A_1$. For the process against which we are discriminating, we will use the respective symbols Q_t, Q_t^*, B^*, B_o, B_1, B.

3.2 Let us first consider the case where our temporal space consists of the non-negative integers.

Theorem 1:

$$\psi(t,x) = \sum_{k=0}^{t-1} P_k^* \, \psi(1,x) \quad . \qquad (3.2.1)$$

Proof: In this case $\phi(1,x) = \psi(1,x)$ then

$$\psi(t,x) = I_{\mathcal{B}_o}\left(\bigvee_{k=1}^{t} \mathcal{B}_k \;\middle|\; x\right)$$

$$= I_{\mathcal{B}_o}\left(\bigvee_{k=1}^{t-1} \mathcal{B}_k \;\middle|\; x\right) + I_{\mathcal{B}_o}\left(\mathcal{B}_t \;\middle|\; \bigvee_{k=1}^{t-1} \mathcal{B}_k , x\right) \quad . \qquad (3.2.2)$$

By the Markov property, we have

$$I_{\mathcal{B}_o}\left(\mathcal{B}_t \;\middle|\; \bigvee_{k=1}^{t-1} \mathcal{B}_k , x\right) = I_{\mathcal{B}_o}(\mathcal{B}_t | \mathcal{B}_{t-1}, x) = P_{t-1}^* \, \psi(1,x) \quad . \qquad (3.2.3)$$

Thus

$$\psi(t,x) = \psi(t-1,x) + P^*_{t-1}\,\psi(1,x) \quad . \tag{3.2.4}$$

(3.2.1) follows by induction.

Corollary 1: If $\psi(1,\cdot)$ is constant, then we have simply (since the P^*_k are transition operators)

$$\psi(t,x) = t\,\psi(1,x) \quad . \tag{3.2.5}$$

3.3 Now let us consider the continuous time case. Denote by $\mathcal{G}_{t,n}$ the algebra $\bigvee\limits_{k=1}^{n} \mathcal{B}_{\frac{kt}{n}}$. The process restricted to this algebra can be considered a discrete-time process, with transition operator $P^*_{t/n}$. Hence we have, from (3.2.4)

$$I_{\mathcal{B}_o}(\mathcal{G}_{t,n}|x) = \sum_{k=0}^{n-1} P^*_{\frac{kt}{n}}\,\phi(\tfrac{t}{n}, x) \quad . \tag{3.3.1}$$

We have that $\bigvee\limits_{n=1}^{\infty} \mathcal{G}_{t,n} = \mathcal{C}_t$ by our separability assumption, and hence

$$\lim_{n\to\infty} I_{\mathcal{B}_o}(\mathcal{G}_{t,n}\,|\,x) = \psi(t,x) \quad . \tag{3.3.2}$$

Let us denote the operator $\frac{1}{n}\sum\limits_{k=0}^{n-1} P^*_{\frac{kt}{n}}$ by $S_{n,t}$. Then (3.3.1) becomes

$$I_{\mathcal{B}_o}(\mathcal{G}_{t,n}\,|\,x) = n\,S_{n,t}\,\phi(\tfrac{t}{n}, x) \quad . \tag{3.3.3}$$

3.4 Before we proceed, let us define a concept of <u>infinitesimal information</u>, which we will denote by L, by

$$L(x) = \lim_{t \downarrow 0} \frac{\phi(t,x)}{t} \tag{3.4.1}$$

whenever the limit exists.

Now, the operator $S_{n,t}$ converges, (since the semi-group is continuous) to

$$\frac{1}{t}\int_0^t P_s^* \, ds \quad .$$

Hence, as all the operators are continuous, we find that

$$I_{\mathcal{B}_o}(\mathcal{C}_{o,t} | x) = \psi(t,x) = \lim_{n\to\infty} (\mathcal{G}_{t,n} | x)$$

$$= \lim_{n\to\infty} S_{n,t} \; n \, \phi(\frac{t}{n} , x)$$

$$= \int_0^t P_s^* \, ds \, L(x) \quad , \tag{3.4.2}$$

provided that L is defined as a <u>uniform</u> limit.

We enunciate the result as

<u>Theorem 2</u>: If L, the infinitesimal information, is defined as a uniform limit, then

$$\psi(t,\cdot) = \int_0^t P_s^* \, ds \, L \quad . \tag{3.4.3}$$

3.5 Suppose that we have that the function $\phi(t,\cdot)$ is constant for every t. This will be true if the following condition is satisfied:

For every $t \varepsilon T$ and every $x,x' \varepsilon E$, there exists a (measurable) permuation π of E such that

$$\begin{aligned} P_t(x,\pi E) &= P_t(x',E) \\ Q_t(x,\pi E) &= Q_t(x',E) \quad . \end{aligned} \tag{3.5.1}$$

The aforementioned consequence of this condition is easily verifield from the fact that a non-singular transformation leaves the information invariant [29, Chapter 2, Corr. 4.1].

Condition (3.5.1) is satisfied in particular if the state space has a group structure, and the process has independent increments. In this case (3.4.3) takes on a very simple form:

Corollary: If (3.5.1) is satisfied then

$$\psi(t,x) = Lt \quad . \tag{3.5.2}$$

3.6 We now have a theorem which ensures the existence of the infinitesimal information.

Theorem 3: If $\lim\limits_{t\downarrow 0} P_t(x,\{x\}) = \lim\limits_{t\downarrow 0} Q_t(x,\{x\}) = 1$ and if both $\frac{P_t(x,F)}{t}$ and $\frac{Q_t(x,F)}{t}$ converge uniformly in $F \varepsilon \{x\}^c \mathcal{E}$ for every $x \varepsilon E$, then

the infinitesimal information $L(x) = \lim_{t \downarrow 0} \frac{\phi(t,x)}{t}$ exists and is equal to $A(x,\{x\}) - B(x,\{x\}) + I[A_1(x,\cdot),B_1(x,\cdot)]$.

Proof: We have that

$$\phi(t,x) = \sup_{A \subset E} \sum P_t(x,F) \log \frac{P_t(x,F)}{Q_t(x,F)} \quad , \quad A \text{ finite.}$$

The summation extending over all atoms F of A .

Because of the separability of E , we have in fact a sequence $A_n \uparrow E$, such that

$$\phi(t,x) = \lim_{n \to \infty} I(P_t(x,\cdot) \ , \ Q_t(x,\cdot) \ ; \ A_n) \quad .$$

Without loss of generality, we may assume that $\{x\} \varepsilon A_n \ \forall \ n$. Then we have

$$\phi(t,x) = \lim_{n \to \infty} I(P_t(x,\cdot),Q_t(x,\cdot);\{x\}^c A_n) + I(P_t(x,\cdot),Q_t(x,\cdot);\{\{x\}\}$$

$$= \phi_1(t,x) + \phi_o(t,x) \quad . \tag{3.6.1}$$

Now

$$\lim_{t \downarrow 0} t^{-1} I(P_t(x,\cdot),Q_t(x,\cdot);\{x\}^c A_n)$$

$$= \lim_{t \downarrow 0} \sum \frac{P_t(x,F)}{t} \log ((\frac{P_t(x,F)}{t})/ (\frac{Q_t(x,F)}{t}))$$

the summation extending over all atoms F of $\{x\}^c A_n$

$$= I(A_1(x,\cdot),B_1(x,\cdot);A_n) \tag{3.6.2}$$

and because the convergence is uniform in the elements of $\{x\}^c$ E we may interchange limits, so that

$$\lim_{t\downarrow 0} t^{-1} \phi_1(t,x) = \lim_{n\to\infty} I(A_1,(x,\cdot),B_1(x,\cdot);A_n)$$

$$= I[A_1(x,\cdot),B_1(x,\cdot)] \quad . \tag{3.6.3}$$

It follows readily from l'Hospital's rule that

$$\lim_{t\downarrow 0} t^{-1} \phi_o(t,x) = A(x,\{x\}) - B(x,\{x\}) \quad . \tag{3.6.4}$$

Hence our theorem is proved.

3.7 Equation (3.4.3), though interesting, is not very useful, because its right side could be rather cumbersome to evaluate. We therefore present another theorem.

Theorem 4: If (3.4.1) holds uniformly in x and if $L \varepsilon \mathcal{C}$, the domain of A^* then

$$\left(\frac{\partial}{\partial t} - A^*\right)\psi = L \quad . \tag{3.7.1}$$

Proof: From (3.4.3) and (1.4.8) we have

$$A^* \psi(t,\cdot) = (P_t^* - 1)L \quad . \tag{3.7.2}$$

From (3.4.3) we also see that

$$\frac{\partial\psi}{\partial t} = P_t^* L \quad . \tag{3.7.3}$$

Combining (3.7.2) and (3.7.3) we obtain (3.7.1).

Corollary: If A_o and B_o are bounded and (1.5.3) holds uniformly in x, then (3.7.1) holds, where L is given by Theorem 3.

REFERENCES

[1] Aczel, J. "On different characterization of entropies", Probability and Information Theory. M. Behara et al (ed.). Springer-Verlag, Berlin 1969, pp. 1-11.

[2] Birkhoff, Garrett. Lattice Theory (2nd edition). American Mathematical Society, New York, 1948.

[3] Black, Max. "Probability", The Margins of Precision, Cornell University Press, Ithaca, 1970; reprinted from The Encyclopedia of Philosophy, Paul Edwards (ed.), Vol. 4, MacMillan, New York 1967, pp. 169-181.

[4] Carathéodory, C. "Entwurf für eine Algebraisierung des Integralbegriffs", S.B. bayer. Akad. Wiss (1938), pp. 27-69.

[5] Cramer, Harold. Mathematical Methods of Statistics, Princeton University Press, Princeton, 1946.

[6] Csiszár, Imre. "Eine informationtheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten", A Magar Tudmányos Akademia Matematikai Kutató Intezetének Közleményi, 8 (1963), pp. 85-108.

[7] Csiszár, Imre. "Information type measures of difference of probability distributions and indirect observations", Studia Scientiarum Mathematicum Hungarica 2(1967), pp. 299-318.

[8] Csiszár, Imre. "On topological properties of f-divergences", ibid. pp. 329-339.

[9] Doob, J.L. Stochastic Processes. Wiley, New York, 1953.

[10] Dynkin, E.B. Markov Processes, Academic Press, New York; Springer-Verlag, Berlin, 1965.

[11] Feinstein, Amiel. Foundations of Information Theory, McGraw-Hill, New York, 1958.

[12] Feller, William. An Introduction to Probability Theory and its Applications, Vol. 2, 2nd edition, Wiley, New York, 1971.

[13] Fisher, R.A. "On the mathematical foundation of theoretical statistics". Contributions to Mathematical Statistics, Wiley, New York 1950, reprinted from Philosophical Transactions of the Royal Society of London, 222A (1922), pp. 309-368.

[14] Fisher, R.A. "Theory of statistical estimation", op. cit., reprinted from Proceedings of the Cambridge Philosophical Society, 22, p. 5, pp. (1925) pp. 700-725.

[15] Fisher, R.A. "The logic of inductive inference", op. cit., reprinted from Journal of the Royal Statistical Society, 98 pt. 1 (1935) pp. 39-54.

[16] Fréchet, Maurice. "Sur la notion de voisinage dans les ensembles abstraits", Bulletin des Sciences Mathematiques 42 (1918), pp. 138-156.

[17] Fréchet, Maurice. Les espaces abstraits. Gauthier-Villars, Paris, 1928.

[18] Ghurye, S.G. "Information and sufficient subfields", Annals of Mathematical Statistics, 39 (1968), pp. 2056-2066.

[19] Halmos, Paul R. "The foundations of probability", American Mathematical Monthly, 51 (1944) pp. 497-510.

[20] Halmos, Paul R. Measure Theory. Van Norstrand, Princeton, 1950.

[21] Hille, Einar. Functional Analysis and Semi-Groups, American Mathematical Society, 1948.

[22] Ingarden, R.S., and Urbanik, K. "Information without probability", Colloquium Mathematican, 9 (1962).

[23] Ingarden, R.S. "Simplified axioms for information without probability", Prace Matematyczne, 9 (1965), pp. 273-282.

[24] Kampé de Fériet, Joseph et Forte, Bruno. "Information et probabilité", Comptes Rendus de l'Academic des Sciences de Paris, 265 A (1967).

[25] Kappos, Demetrios, A. Strukturtheorie der Wahrscheinlichkeitsfelder und raume, Springer-Verlag, Berlin (1960).

[26] Kappos, Demetrios, A. Probability Algebras and Stochastic Spaces, Academic Press, New York (1969).

[27] Kolmogorov, A.N., Gel'fand, I.M., and Yaglom, A.M. "On the general definition of the amount of information" Doklady Akademii Nauk SSSR, 111 (1956), pp. 745-748.

[28] Kullback, S., and Leibler, S. "Information and sufficiency", Annals of Mathematical Statistics, 22 (1951), pp. 79-86.

[29] Kullback, S. Information Theory and Statistics. Dover, New York, 1968.

[30] Lee, P.M. "On the axioms of information theory", Annals of Mathematical Statistics 35 (1964), pp. 414-441.

[31] Lindley, D.V. "On a measure of the information provided by an experiment", ibid. 27 (1956), pp. 986-1005.

[32] Meyer, Paul-André. Probability and Potentials, Blaisdell, Waltham 1966, translation of Probabilités et potentiel, Hermann, Paris, 1966.

[33] Meyer, Paul-André. Processus de Markov, Springer-Verlag, Berlin, (1967).

[34] Parry, William, Entropy and Generators in Ergodic Theory, Benjamin, New York (1969).

[35] Rao, C. Rhadhakrishna. "Information and accuracy obtainable in an estimation of a statistical parameter", Bulletin of the Calcutta Mathematical Society, 37 (1945), pp. 81- .

[36] Rényi, Alfred. "On measures of entropy and information". Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1, University of California Press, Berkeley, 1961, pp. 547-561.

[37] Royden, H.L. Real Analysis (2nd edition), Macmillan, New York, 1968.

[38] Sakaguchi, Minoru. Information and Decision Making, George Washington University, Washington, 1964.

[39] Shannon, Chaude E. The Mathematical Theory of Communication, (with an introductory article by Warren Weaver) University of Illinois Press, Urbana 1949, reprinted from Bell System Technical Journal.

[40] Wiener, Norbert. Cybernetics (2nd edition). MIT Press, Cambridge, Mass. 1961.

APPENDIX 1

A convergence theorem:

Let $(\Omega,\mathcal{F},P)$ be a probability space.

Let $\{\mathcal{A}_n\}$ be a family of σ-algebras increasing to $\mathcal{A} \subset \mathcal{F}$.

Let f be a function convex on $[0,\infty)$, and let p be a non-negative integrable function.

Let $\phi_n = \mathcal{E}(p|\mathcal{A}_n)$ and $\phi_\infty = \mathcal{E}(p|\mathcal{A})$. Further, let $i_n = f \circ \phi_n$.

Then $i_n \to i_\infty$ and $\mathcal{E}(i_n) \to \mathcal{E}(i)$.

Proof: $\{\phi_n \; ; \; n = 1,\dots,\infty\}$ is a martingale, and hence $\phi_n \to \phi_\infty$ [28]. As convexity implies continuity; $i_n \to i_\infty$, $\{i_n \; ; \; n = 1,2,\dots\}$ forms a sub-martingale. Hence $n < m \Longrightarrow$

$$\mathcal{E}(i_n) \leq \mathcal{E}(i_m) \leq \mathcal{E}(i_\infty) \quad .$$

Thus $\mathcal{E}(i_n)$ is increasing and bounded above, and converges to some number $\leq \mathcal{E}(i_\infty)$. Now $\mathcal{E}(i_n) = \mathcal{E}(i_n^+) - \mathcal{E}(i_n^-)$.

By Fatou's lemma,

$$\lim_n \; \mathcal{E}(i_n^+) \geq \mathcal{E}(i_\infty^+)$$

so it only remains to prove that

$$\lim_n \; \mathcal{E}(i_n^-) = \mathcal{E}(i_\infty^-) \quad .$$

Let x_o be the largest number such that $f(x_o) = 0$. If there is no such number, then $f > 0$, $i_n^- = 0$ and the theorem is proved. Otherwise, let $\alpha x + \beta$ be the line of support of f at x_o. i.e.,

$$\alpha x_o + \beta = 0 \quad \text{and} \quad f(x) \geq \alpha x + \beta \quad .$$

<u>Case 1</u>: Suppose $\alpha > 0$. Let

$$\begin{aligned} g(x) &= -\alpha x - \beta \quad &&\text{for} \quad x < x_o \\ &= 0 &&\text{for} \quad x \geq x_o \ . \end{aligned}$$

Then g is convex, bounded and integrable.

$$i_n^- = f^- \circ \phi_n \leq g \circ \phi_n$$

hence by bounded convergence $E(i_n^-) \to E(i_\infty^-)$.

<u>Case 2</u>: Suppose $\alpha < 0$. Let

$$\begin{aligned} g(x) &= -\alpha x - \beta \quad &&\text{for} \quad x > x_o \\ &= 0 &&\text{for} \quad x < x_o \ . \end{aligned}$$

Then $f^- \leq g$

$$i_n^- = f^- \circ \phi_n \leq g \circ \phi_n \leq \phi_n \quad .$$

Hence, by dominated convergence $E(i_n^-) \to E(i_\infty^-)$.

APPENDIX 2

A semi-group theorem:

Let $(E,\mathcal{E})$ be a measurable space.

Let $\{P_t\}$ be a semi-group of Markovian kernels on $(E,\mathcal{E})$ and let $\{P_t^*\}$ be the associated linear operators.

Suppose that

$$A_o(x,\ x\ ) = \lim_{t\downarrow 0} \frac{P_t(x,\{x\}) - 1}{t}$$

and

$$A_1(x,F) = \lim_{t\downarrow 0} \frac{P_t(x,F)}{t}$$

exist uniformly in x and $F \in \{x\}^c\ \mathcal{E}$. Then the limit $\frac{P_t^* - I}{t} = A^*$ exists uniformly as $t \downarrow 0$ and $A = A_o + A_1$ is the kernel of this infinitesimal operator.

Proof:

$$\left|\left|\frac{P_t^* - I}{t} - A^*\right|\right| = \sup_{||f||=1}\ \sup_{x} \left|\frac{\int P_t(x,dy)\ f(y)-f(x)}{t} - \int A(x,dy)f(y)\right| .$$

Let us separate the integral into two parts: over $\{x\}$ and over $E - \{x\}$. Of course

$$\left| \frac{\int P_t(x,dy)\ f(y)-f(x)}{t} - \int A(x,dy)\ f(y) \right|$$

$$\leq \left| \frac{\int_{\{x\}} P_t(x,dy)\ f(y)-f(x)}{t} - \int_{\{x\}} A(x,dy) f(y) \right|$$

$$+ \left| \int_{E-\{x\}} \frac{P_t(x,dy) f(y)}{t} - \int_{E-\{x\}} A(x,dy) f(y) \right| \quad .$$

Now, the first quantity in $| \ |$ is merely

$$\left| \frac{P_t(x,\{x\}) - 1}{t} - A(x,\{x\}) \right| \ \left| f(x) \right|$$

and as we are taking the sup over functions for which $||f|| = 1$, $|f(x)| \leq 1$. Thus we can make the first quantity $< \varepsilon_1$. The second quantity in $| \ |$ is less than or equal to

$$\int_{E-\{x\}} \left| \frac{P_t(x,dy)}{t} - A(x,dy) \right| \ \left| f(y) \right|$$

and again as $|f(y)| \leq 1$ we have the integral $\leq$

$$\sup_{F \varepsilon \{x\}^c \mathcal{E}} \left| \frac{P_t(x,E-\{x\})}{t} - A(x,E-\{x\}) \right| \leq \varepsilon_2 \quad ,$$

because the convergence is uniform in both x and F. Thus the theorem is proved.